

Análise de Super-Resolução por Aprendizado Profundo sob Diferentes Degraadações

Laura Naomi Seto, Leonardo Oliveira

CCGT - Departamento de Computação

Universidade Federal de São Carlos

Sorocaba, Brasil

{ laura.naomi, leonardooliveirapedro } @estudante.ufscar.br

Resumo—O *upscaleing* de imagens tem como objetivo aumentar a resolução de uma imagem digital. Explorando soluções de modelos baseados em aprendizados profundo com CNNs, GANs e Transformers treinados, o projeto analisa como cada um se comporta quando apresentado a imagens de baixa resolução com diferentes degradações.

I. INTRODUÇÃO

O *upscaleing* de imagens, também referido como Super Resolução (SR) de imagens, é um problema de processamento de imagens que tem como objetivo aumentar a resolução de uma imagem digital expandindo suas dimensões a fim de contribuir para a melhora na sua percepção visual. Nota-se que o processo não é somente uma tarefa final com a intenção já mencionada, ele também atua como ferramenta de apoio em múltiplas tarefas de visão computacional quando há limitações de resolução nos dados de entrada.

Segundo [1], as estratégias de abordagem podem ser classificadas em quatro grupos principais: baseadas em interpolação espacial com soluções como vizinhança mais próxima, interpolação bilinear e bicúbica e por modelos baseados em filtros (convoluções); baseadas em reconstrução ou regularização que exigem múltiplas imagens de entrada; baseados em aprendizado profundo com soluções de Redes Neurais Convolucionais (CNNs) e Redes Adversariais Generativas (GANs); e baseados em Transformers.

As técnicas devem lidar com desafios como a perda de detalhes - quando as dimensões são expandidas, mas o preenchimento realizado pode não ser fidedigno a imagem original; o borramento - gerado em imagens com bordas borradadas e menos nítidas; e ambiguidade ou problema inverso mal posto, já que muitas imagens em alta resolução poderiam gerar uma mesma versão em baixa resolução e por isso não há uma única resposta correta para o aumento da resolução.

O projeto propõe uma análise comparativa sistemática sobre as abordagens baseadas em aprendizado profundo para SR, executando modelos pré-treinados de diferentes arquiteturas para investigar seu desempenho com base em métricas objetivas e subjetivas sob diversos cenários de degradação, identificando a solução mais robusta entre as investigadas.

II. TRABALHOS RELACIONADOS

Diversas estratégias de aprendizado profundo baseadas em Redes Neurais Residuais foram encontrados na literatura para

SR.

A geração de abordagens que se baseiam em CNN desenvolveu soluções com foco na otimização de métricas objetivas como o PSNR. Dentre eles, o EDSR (*Enhanced Deep Super-Resolution Network*) [2], que venceu o desafio NTIRE 2017, é uma rede obtida a partir de adaptações da arquitetura do SRResNet. O trabalho removeu camadas BN, aumentou a arquitetura, introduziu o escalonamento residual para estabilizar o treinamento e é considerado um *baseline* de PSNR.

Diferentemente das CNNs, os modelos com GANs têm foco na qualidade perceptual do *upscalling*. O SRGAN (Super Resolution GAN) [3] foi o primeiro a utilizar uma SRResNet como gerador em uma GAN no contexto de SR para obter uma função de perda perceptual baseada na diferença entre os *feature maps* extraídas de uma rede VGG pré-treinada. Seu sucessor, ESRGAN (Enhanced SRGAN) [4], que venceu o desafio PIRM2018-SR, aprimora o SRGAN substituindo blocos residuais por RRDBs, removendo camadas BN, utilizando *perceptual loss* com *features* antes da ReLU e adotando *relativistic GAN* como discriminador.

Por fim, a abordagem com arquitetura *Transformers* inova ao capturar relações contextuais complexas com mecanismos de auto-atenção e é utilizada tanto para a chamada SR clássica, como para as tarefas perceptuais. O SwinIR (*Swim Transformer for Image Restoration*) [7] utiliza o mecanismo de atenção do *Swin Transformer* para modelar dependências de longo alcance e superar a limitação do campo receptivo local das convoluções padrão.

III. CONJUNTO DE DADOS

Para realizar a avaliação e aferir o desempenho dos modelos selecionados, alguns *datasets benchmark* de Super Resolução foram considerados. O Set14 é um conjuntos de 14 imagens que contém imagens consideradas clássicas no problema de *upscaleing* e são utilizados para testes rápidos. O BSD100 é uma coleção de 100 imagens que cobrem conteúdos como natureza, pessoas, objetos e cenas gerais, tornando-o ideal para medir a capacidade de generalização dos modelos. O Urban100 contém 100 imagens de alta resolução com paisagens urbanas, compostas por estruturas repetitivas e texturas complexas com detalhes geométricos com bordas nítidas e padrões regulares. O DIV2K é um conjunto muito popular para treinar modelos de SR, porque oferece 1000 imagens (particionadas

em 800 imagens de treino, 100 de validação e 100 de teste) de conteúdo variado com resolução de 2k. Além disso, foram utilizadas imagens coletadas do website Fancaps.net [9], uma plataforma colaborativa onde fãs compartilham screenshots de filmes, séries e animações e do portal oficial da *Library of Congress* [8] (Biblioteca do Congresso dos EUA).

IV. MÉTODOS

A. Modelos

Como o foco do projeto está na execução de modelos pré-treinados representativos de diferentes gerações de SR baseadas em CNNs, GANs e *Transformers*, avaliando seu desempenho com métricas objetivas e análises subjetivas em cenários diversos de degradação de imagens, os modelos EDSR (CNNs), ESRGAN (GAN) e SwinIR (Transformers) foram escolhidos para a análise.

- **EDSR (CNNs):**

- Fonte: eugeniesow/edsr-base [Hugging Face]
- Dataset de treinamento: DIV2K ampliado com técnicas de data augmentation
- Downsample de treino: Interpolação bicúbica com redução por escala x2, x3 e x4

- **ESRGAN (GANs):**

- Modelo: RRDB_ESRGAN_x4.pth
- Fonte: xinntao/ESRGAN [GitHub]
- Dataset de treinamento: DIV2K, Flickr2K e OutdoorsSceneTraining
- Downsample de treino: Interpolação bicúbica com redução por escala x4

- **SwinIR (Transformer):**

- Modelo: 001_classicalSR_DIV2K_s48w8_SwinIR-M_x4.pth
- Fonte: JingyunLiang/SwinIR [GitHub]
- Dataset de treinamento: DIV2K
- Downsample de treino: Não explicitado no artigo, mas muito provavelmente por interpolação bicúbica (degradação padrão do DIV2K)

Essa seleção permite comparar arquiteturas distintas em condições controladas, já que todos foram originalmente treinados com degradação bicúbica, garantindo fidelidade nas métricas tradicionais.

B. Degradações

Para criar as entradas de baixa resolução que dão cobertura para diferentes cenários, as imagens dos *datasets* utilizadas na avaliação foram submetidas a processos de degradação com *downsamplings* na escala x4 e posteriormente a borramentos e adições de ruídos.

Os *downsamplings* utilizados são aqueles disponíveis na biblioteca Pillow:

- **Interpolação Nearest-Neighbor:** Seleciona o valor do pixel mais próximo na imagem original para determinar o valor do novo pixel. Pode produzir resultados com *aliasing* (bordas serrilhadas).

- **Interpolação Bilinear:** Utiliza a média ponderada dos quatro pixels mais próximos da imagem original para calcular o valor do novo pixel. Os pesos são determinados com base na distância de cada vizinho ao pixel a ser preenchido. Produz resultados mais suaves do que o Nearest-Neighbor.

- **Interpolação Bicúbica:** Utiliza a média ponderada dos 16 vizinhos (grade 4x4) mais próximos da imagem original para calcular o valor do novo pixel. Produz resultados mais suaves do que a interpolação Bilinear.

- **Interpolação de Lanczos:** Aplica a função *sinc* normalizada (função seno cardinal normalizada) para ponderar um número de pixels vizinhos e calcular o valor do novo pixel. É um método de interpolação avançado, com desempenho de destaque para *downsamplings* com boa nitidez.

- **Interpolação de Hamming:** Aplica a função *sinc* truncada (versão limitada da função *sinc*) e utiliza uma outra função matemática que modifica o *kernel* de interpolação, chamada de janela de Hamming, para minimizar o truncamento. Apresenta um bom equilíbrio entre nitidez e supressão de distorções de bordas.

Além dos métodos de *downsampling*, aplicam-se borramentos e ruídos para investigar a robustez dos modelos em diversos cenários:

- **Interpolação de Lanczos + Gaussian Blur:** Simulação de desfoque óptico devido a limitações ópticas de câmeras, lentes e movimentos sutis. O modelo deve aprender a reconstruir detalhes finos atenuados pelo desfoque.

- **Interpolação de Lanczos + Motion Blur:** Simulação de desfoque causado por movimento da câmera ou do objeto durante a exposição. Como o borramento não é uniforme em todas as direções, como no caso do Gaussiano, o modelo deve aprender a remover artefatos de desfoque com um padrão específico.

- **Interpolação de Lanczos + Gaussian Noise:** Simula o ruído eletrônico de sensores de câmeras. O modelo deve aprender a diferenciar ruídos aleatórios e detalhes legítimos da imagem.

- **Interpolação de Lanczos + Salt and Pepper Noise:** Simula falhas de sensores, erros de transmissão dos dados e corrupções de arquivos. O modelo deve aprender a recuperar *outliers* extremos, inferindo valores corretos.

Todas as imagens LR geradas estão disponíveis em `projeto/datasets/<dataset>/LR_<downgrade_type>/x4/` para consulta.

C. Métricas

Para medir o desempenho dos modelos empregados no estudo, utilizaremos métricas clássicas (de fidelidade) e perceptuais. Dentre as métricas objetivas principais estão: o RMSE (*Root Mean Square Error*) que calcula a média da diferença quadrática entre os pixels das imagens; o PSNR (*Peak Signal-to-Noise Ratio*) que mede a razão entre a maior potência de

um sinal e a potência do ruído que afeta a sua representação; o SSIM (*Structural Similarity Index*) que analisa a mudança percebida na informação estrutural das imagens, considerando a luminância e o contraste; e o LPIPS (*Learned Perceptual Image Patch Similarity*) que afere a similaridade perceptual com base em *features* extraídas de redes pré-treinadas. Ademais, realizaremos uma análise visual direta comparando imagens originais a aquelas geradas pelos modelos de SR, sem empregar métricas qualitativas específicas.

Após a criação das imagens LR e a execução do *upsampling* com os modelos selecionados, as métricas RMSE, PSNR, SSIM e LPIPS serão calculadas com base nas imagens originais (HR) e aquelas reconstruídas. Ademais, realizaremos uma comparação visual direta para obter noções gerais sobre a qualidade perceptual dos resultados.

V. EXPERIMENTOS E RESULTADOS

Para a validação quantitativa, os *datasets* Set14, BSD100, URBAN100 e DIV2K, cada um limitado a 15 imagens, foram utilizados. Essa restrição foi aplicada pela moderação dos recursos computacionais, tendo em vista que esses conjuntos de imagens foram submetidos a 9 diferentes degradações que geraram, no final, 540 imagens. Para a análise qualitativa, além daqueles já citados, os *datasets* fancaps.net e loc.gov foram incorporados para abordar cenários diversificados.

O fluxo de trabalho consistiu em: (i) *Downsampling* das imagens por interpolações diversas (escala $\times 4$) e aplicação ou não de borramentos e ruídos; (ii) Carregamento dos modelos pré-treinados e *upsampling*; e (iii) Cálculo das métricas objetivas/Análise subjetiva.

Os *scripts*, os modelos e os *datasets* utilizados para a execução de cada etapa, assim como as imagens geradas (*downsamplings* e SRs) podem ser encontrados diretamente no diretório geral do projeto original disponibilizado aqui.

A. Análise quantitativa

Analizando o RMSE com a Figura 1, que reflete a diferença direta entre os pixels das imagens, verificamos, para a maioria das degradações, uma diminuição de pelo menos 0.25 nos erros médios aferidos, fato que reflete uma melhora na imagem em termos de fidelidade direta. Como exceção, estão as degradações com Interpolação *Nearest-Neighbor*, Interpolação de Lanczos+*Gaussian Noise* e a Interpolação de Lanczos+*Salt and Pepper Noise*. Nesta métrica, os modelos EDSR e SwinIR detêm os melhores desempenhos de forma consistente, ficando um a frente do outro e vice e versa em diferentes cenários.

Sobre a métrica PSNR na Figura 2, os modelos EDSR e SwinIR superam as imagens LR de forma consistente, indicando uma redução significativa nas distorções das imagens. Sob o ponto de vista das degradações, é possível observar que a Interpolação de Lanczos+*Gaussian Noise* foi o mais desafiador para todos os modelos.

O SSIM, na Figura 3, que afere a qualidade da imagem como uma mudança percebida na informação estrutural, apresentou tendências similares ao PSNR. Mais uma vez, o EDSR e o SwinIR tomaram a liderança com médias acima de 0.7

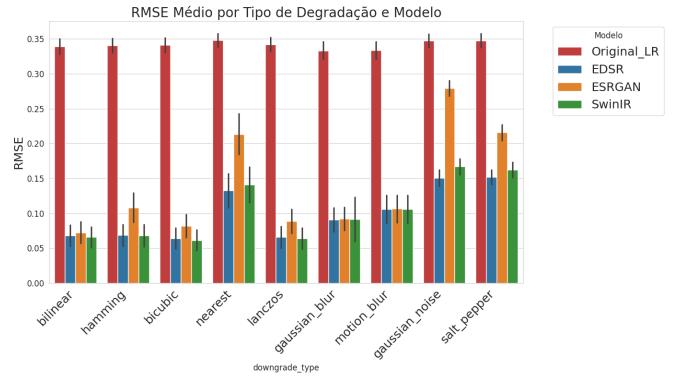


Figura 1: RMSE médio por tipo de degradação e modelo.

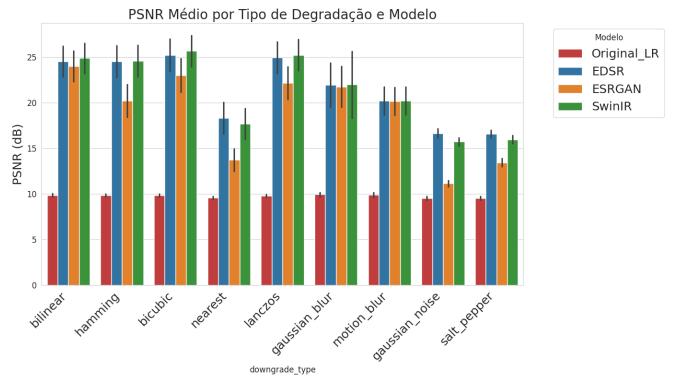


Figura 2: PSNR médio por tipo de degradação e modelo.

para degradações mais leves. Aqui, é interessante apontar que, além da média significativamente baixa de todos os modelos na Interpolação de Lanczos+*Gaussian Noise*, a Interpolação *Nearest-Neighbor* e a Interpolação de Lanczos+*Salt and Pepper Noise* foram particularmente difíceis para o ESRGAN.

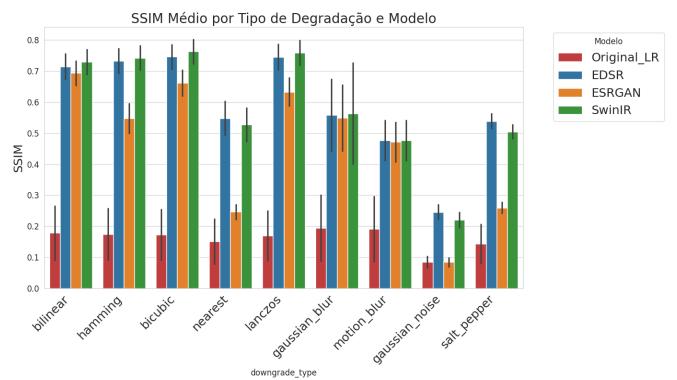


Figura 3: SSIM médio por tipo de degradação e modelo.

Para a métrica LPIPS, a Figura 4 mostra que o ESRGAN têm médias LPIPS mais baixas, indicando melhor desempenho em relação aos modelos anteriores em todas as *downgrades*, com exceção daqueles em que ele já apresentou dificuldade expressiva anteriormente. Um apontamento importante é que

para a Interpolação de Lanczos+*Gaussian Noise*, o ESRGAN superou o LPIPS da LR original, sugerindo que o modelo introduziu artefatos ou reconstruções que são tidos como uma piora da imagem LR em termos perceptuais.

Considerando que o EDSR é um modelo otimizado para PSNR e SSIM com função de perda de reconstrução pixel a pixel, seu desempenho inferior ao ESRGAN em uma métrica perceptual é compreensível, já que o segundo foi projetado para priorizar o realismo visual sobre a fidelidade direta. Acerca do SwinIR, apesar de não ser exclusivamente um modelo perceptual como o ESRGAN, a expectativa, não atendida, era de que sua arquitetura permitisse capturar detalhes e texturas de forma mais eficaz, ainda que não tenha sido treinado para as situações apresentadas aqui.

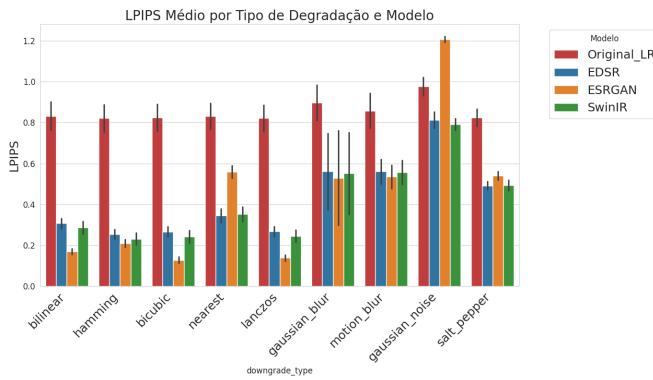


Figura 4: LPIPS médio por tipo de degradação e modelo.

Como mostra a Figura 5, no geral, os modelos SwinIR e EDSR têm um desempenho superior nas métricas de fidelidade (PSNR) e similaridade estrutural (SSIM), considerando a média em todos os tipos de degradações aplicadas e os variados *datasets*. Além disso, os mesmos modelos apresentam um erro menor (RMSE) e uma melhora razoável na qualidade perceptual (LPIPS) como expresso na Figura 6. Essas afirmações podem ainda ser confirmadas na Tabela I.

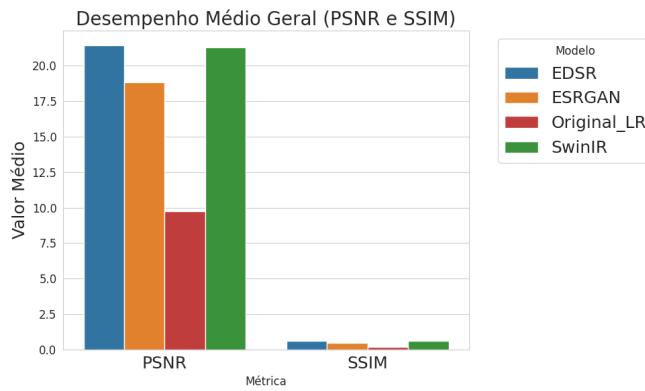


Figura 5: Desempenho médio geral nas métricas PSNR e SSIM (valores altos indicam melhores desempenhos).

Em um apanhado geral, partindo de uma análise que avalia a robustez e a estabilidade dos modelos sob diver-

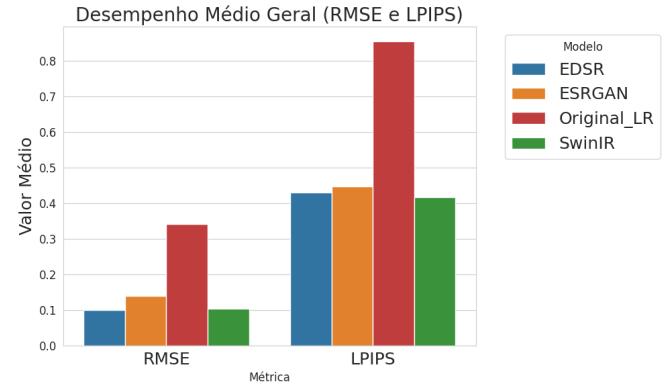


Figura 6: Desempenho médio geral nas métricas RMSE e LPIPS (valores baixos indicam melhores desempenhos).

sas degradações considerando que eles foram treinados em condições similares, o SwinIR é o modelo mais robusto e completo, com resultados consistentes em todas as métricas, apresentando uma liderança quase incontestável, seguido pelo EDSR que deve ocupar o segundo lugar.

Modelo/LR	RMSE	PSNR	SSIM	LPIPS
LRs geradas	0.3415	9.74 dB	0.1621	0.8543
EDSR	0.0999	21.41 dB	0.5892	0.4296
ESRGAN	0.139864	18.83 dB	0.4604	0.4464
SwinIR	0.103159	21.30 dB	0.5869	0.4167

Tabela I: Média das métricas por modelo em todos os *datasets* e tipos de degradações.

B. Avaliação subjetiva

1) *Super-Resolução Clássica*: Utilizando os mesmos modelos da análise quantitativa (EDSR, ESRGAN e SwinIR), foi realizada uma avaliação visual para comparar a qualidade perceptual dos resultados.

De modo geral, uma análise inicial das imagens geradas, como visto na Figura 7, revela que os três modelos produzem resultados bastante semelhantes em um primeiro olhar. No entanto, é possível notar que as reconstruções do EDSR apresentam uma nitidez sutilmente inferior em comparação com as do ESRGAN e do SwinIR.



Figura 7: Imagem 038 (BSD100). Ordem: Original, LR (Bicúbico), EDSR, ESRGAN e SwinIR.

Avaliando especificamente a degradação por interpolação bicúbica — o método de *downgrade* com o qual os três modelos foram originalmente treinados —, as diferenças aqui são as menos perceptíveis (Figura 8). A alta similaridade dos

resultados se explica pelo fato de os modelos serem especialistas em reverter essa degradação específica. Ainda assim, o EDSR se mantém como o modelo levemente menos nítido. Essa característica se justifica pelo seu foco primário em otimizar o PSNR em seu treinamento e também devido à sua arquitetura de campo receptivo local (CNN), não criando detalhes que, embora perceptualmente agradáveis (como texturas acentuadas), podem não estar presentes na imagem original de alta resolução, sendo, portanto, considerados "erros" e evitados.



Figura 8: Imagem 098 (URBAN100). Ordem: Original, LR (Bicúbico), EDSR, ESRGAN e SwinIR. Nota-se que o EDSR reconstrói as janelas superiores com a geometria distorcida ("tortas"), um artefato que tanto o ESRGAN quanto o SwinIR corrigem, restaurando as linhas retas. Ambos também adicionam detalhes das esquadrias, não visíveis na versão LR. Entre eles, o SwinIR se diferencia ao introduzir uma textura mais forte na parede à esquerda e mais detalhes na segunda janela.

Em contraste, tanto o ESRGAN quanto o SwinIR geram detalhes com maior contraste e acutância (nitidez percebida) em comparação com o EDSR. Ao analisar apenas estes dois, os resultados para a degradação bicúbica são extremamente parecidos, a tal ponto que as diferenças entre eles só são perceptíveis sob alta ampliação, como observado.

As demais diferenças entre eles se manifestam na reconstrução de detalhes finos e específicos, como em padrões de texturas ou símbolos pequenos presentes nas imagens, conforme pode ser observado na Figura 9 e 10.

A capacidade dos modelos em reconstruir textos e letreiros pequenos, garantindo sua legibilidade, também se mostrou um importante fator de comparação, como ilustrado na Figura 11.

Com base nestas observações, para o cenário de *downgrade* bicúbico, os três modelos produzem resultados visualmente similares, mas com prioridades de reconstrução distintas.

O ESRGAN se destaca ao gerar texturas mais ricas e realistas, focando no realismo perceptual.

O SwinIR, por sua vez, curiosamente demonstra uma melhor performance na manutenção da integridade de estruturas geométricas, como observado nas janelas, mas também mantém maior fidelidade a artefatos da degradação. Apesar disso, a semelhança entre ambos é tão alta que suas diferenças são muito sutis.

Já o EDSR gera o resultado mais suave, cujo desempenho reflete sua otimização para a fidelidade matemática (PSNR) em detrimento de detalhes visuais. O SwinIR prova ser superior ao EDSR pois sua arquitetura Transformer permite uma reconstrução mais fiel de estruturas complexas e texturas, mesmo ambos buscando a melhor pontuação de PSNR.

Contudo, quando o critério de avaliação é a qualidade perceptual, o ESRGAN se sobressai, entregando o resultado



Figura 9: Imagem barbara (Set14). Ordem: EDSR, ESRGAN e SwinIR. O padrão do caderno, distorcido para linhas diagonais pelo *downgrade*, evidencia a diferença entre os modelos. O SwinIR mantém parte do artefato, tornando as linhas diagonais nítidas; o EDSR também as mantém, porém com menos nitidez. Em contraste, o ESRGAN corrige a geometria e restaura as linhas retas, um resultado de seu treinamento adversarial que prioriza o realismo sobre a fidelidade ao artefato de entrada. Nota-se ainda diferenças sutis na textura da toalha de mesa.



Figura 10: Imagem 0900 (DIV2K). Ordem: EDSR, ESRGAN e SwinIR. As diferenças tornam-se mais evidentes na reconstrução de algumas texturas, no entanto, tratam-se de distinções quase imperceptíveis, onde ESRGAN e SwinIR superam o EDSR. É possível notar, nesse caso, que o ESRGAN adiciona detalhes de textura no rádio, enquanto o SwinIR define diferentes símbolos superiores na caixa verde. Ambos se destacam também ao melhorar a qualidade dos detalhes do papel na parede, no entanto, novamente ocorre uma distinção entre os dois: o SwinIR reproduz as linhas diagonais (artefato do *downgrade*) na parte inferior, enquanto o ESRGAN as corrige para linhas retas, priorizando o realismo.

visualmente mais rico (mesmo que em mínimos detalhes) e convincente.

Ao explorar outros tipos de degradação, o comportamento distinto do ESRGAN torna-se mais evidente. Estranhamente, para um grupo de degradações que inclui Lanczos e os filtros de borramento *Gaussian Blur* e *Motion Blur*, os resultados



Figura 11: Imagem ppt 3 (Set14). Ordem: EDSR, ESRGAN e SwinIR. Neste exemplo, o ESRGAN e o SwinIR se destacam ao preservar a legibilidade do texto maior. No texto menor, embora permaneça ilegível em todos os resultados, ambos trazem mais nitidez às letras em comparação com o EDSR. Entre eles, o SwinIR aparenta reconstruir alguns detalhes de separação adicionais nas formas das letras.

de todos os modelos permanecem consistentes e semelhantes entre si. Em contrapartida, nos *downgrades* Bilinear e, principalmente, Hamming o mecanismo do ESRGAN se desvia, tornando sua abordagem única claramente visível.

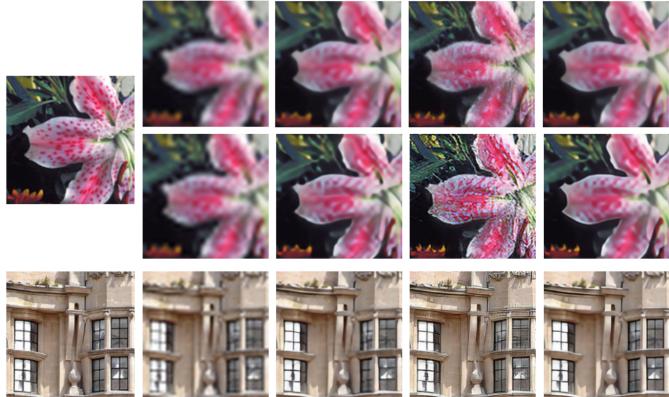


Figura 12: As duas primeiras linhas usam a imagem flowers (Set14), degradada com os filtros Bilinear (linha 1) e Hamming (linha 2). A terceira linha mostra a imagem 053 (URBAN100), também com downgrade Hamming. Para todas as linhas, as colunas são: Original, LR, EDSR, ESRGAN e SwinIR. É notório como o ESRGAN, treinado em bicúbico, reage a degradações desconhecidas de modo destacado. O efeito já é visível na linha 1 (Bilinear), onde ele gera levemente texturas adicionais. Nas linhas 2 e 3 (Hamming), a reação é mais forte: o modelo interpreta mal os artefatos e os transforma em texturas irreais, de forma similar a um filtro de nitidez exagerado.

Embora a imagem possa parecer mais detalhada à primeira vista, essa acutância artificial, na verdade, prejudica a qualidade visual. Tal efeito não representa uma melhora, mas sim

evidencia a abordagem perceptual do ESRGAN.



Figura 13: Imagem 011 (BSD100). Colunas: Original, LR (*Nearest Neighbors*), EDSR, ESRGAN e SwinIR. Neste cenário, o ESRGAN gera detalhes em um padrão mais contido e estruturado, assemelhando-se ao resultado do SwinIR. Isso sugere que o SwinIR, por sua vez, parece interpretar os artefatos "em bloco" dessa degradação como padrões geométricos, e o ESRGAN, neste caso específico, parece seguir uma lógica de reconstrução similar.

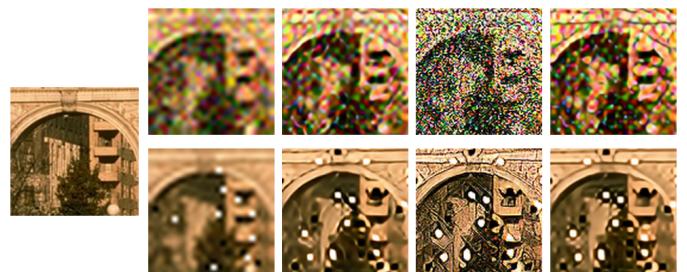


Figura 14: Imagem 022 (BSD100). Linha superior: *Gaussian Noise*. Linha inferior: *Salt and Pepper Noise*. Ordem (colunas): Original, LR, EDSR, ESRGAN e SwinIR. Observa-se que para o Ruído Gaussiano, o ESRGAN parece ter interpretado o ruído fino como detalhes, intensificando sua presença e tornando-o menor e mais potente, o que gerou uma aparente diferença na coloração geral da imagem. No caso do Ruído Salt and Pepper, o ESRGAN demonstra um comportamento curioso, ele gera detalhamentos e texturas por toda a imagem. É importante notar que, como citado, apenas com o *downsampling* Lanczos, o ESRGAN não exibia essa tendência, o que gerou resultados muito parecidos com os dos outros modelos. A presença desse ruído parece ter induzido o modelo a gerar uma quantidade significativa de texturas e contornos.

Conclui-se que, enquanto os modelos EDSR e SwinIR apresentam resultados consistentemente conservadores em todas as degradações devido à sua otimização para PSNR, o ESRGAN exibe um comportamento dual e dependente do tipo de artefato. Apesar de ter sido treinado apenas em bicúbico, ele generaliza de forma surpreendente para degradações como Lanczos e borrões suaves, e de modo um pouco menos conservador com o Nearest Neighbors.

No entanto, degradações que introduzem artefatos de alta frequência específicos, como o "ringing" do filtro Hamming e os pixels impulsivos do ruído Salt and Pepper, expõem sua principal característica: nesses casos, os artefatos são mal interpretados como detalhes, levando o modelo a "alucinar" texturas irrealis em sua busca por realismo perceptual.

Já o filtro Bilinear revelou o limiar de sensibilidade do ESRGAN: mesmo um borrão relativamente simples foi suficiente

para induzir uma leve, porém notável, geração de texturas. Este comportamento é particularmente intrigante ao ser contrastado com os resultados do filtro Lanczos, que não provocou uma reação similar.

2) *Super-Resolução Cega*: Para aprofundar a comparação qualitativa, esta seção introduz três modelos projetados para o desafio da super-resolução cega (*blind SR*). Esses modelos são treinados com degradações complexas e variadas, permitindo-lhes resolver os diversos cenários de degradações abordados aqui, incluindo agora, a compressão JPEG (método com perdas inerente ao próprio formato, que descarta detalhes para reduzir o tamanho do arquivo), algo fortemente presente em imagens de fontes online.

Os modelos são: O Real-ESRGAN [5], que é uma evolução direta do ESRGAN e substitui o simples *downgrade bicúbico* por um modelo de deterioração de "alta ordem", simulando uma cadeia mais realista de problemas (borrão, ruído, redimensionamento, etc.) para melhorar a robustez.

O BSRGAN (Blind-SRGAN) [6], que leva a ideia do Real-ESRGAN um passo adiante, utilizando um modelo de degradação ainda mais complexo e aleatório. Ele embaralha a ordem das distorções aplicadas (*blur*, *downsampling*, *noise*, JPEG) para cobrir um espaço ainda maior de problemas do mundo real.

Finalmente, o SwinIR para diversos tipos de degradações, representado aqui por outro arquivo do modelo: 003_realSR_BSRGAN_DFO_s64w8_SwinIR-M_x4_GAN.pth que eleva a performance ao utilizar a arquitetura Swin Transformer dentro da metodologia de treinamento com múltiplas degradações do BSRGAN. Adicionalmente, uma estrutura GAN é empregada na disputa para garantir que os resultados sejam não apenas robustos, mas também visualmente realistas.

• Real-ESRGAN:

- Modelo: RealESRGAN_x4plus.pth
- Fonte: [xinntao/Real-ESRGAN](#) [GitHub]
- Dataset de treinamento: DIV2K, Flickr2K e OutdoorsTraining

• BSRGAN:

- Modelo: BSRGAN.pth
- Fonte: [cszn/BSRGAN](#) [GitHub]
- Dataset de treinamento: DIV2K, Flickr2K, WED e FFHQ.

A análise dos resultados demonstra a notável eficácia dos modelos de super-resolução cega. Embora os resultados sejam, no geral, muito semelhantes, foram observadas especializações sutis: o BSRGAN, por exemplo, se destaca pela reconstrução mais natural de rostos, um reflexo de seu treinamento com o *dataset* FFHQ. Em contrapartida, o SwinIR demonstrou uma superioridade pontual em cenários específicos, como na remoção *Salt and Pepper Noise* e ao lidar com o *Motion Blur*, onde sua arquitetura Transformer parece oferecer vantagens nessas situações, devido a capacidade de lidar com dependências de longo alcance.



Figura 15: Imagem 020 (URBAN100). Linha superior: Lanczos + Ruído Gaussiano. Linha inferior: Lanczos + Ruído Salt and Pepper. Colunas: LR, SwinIR, BSRGAN e Real-ESRGAN. Nota-se que os ruídos são corrigidos de maneira mais efetiva nessa abordagem. Para o Ruído Gaussiano, o BSRGAN é o mais afetado, exibindo uma leve alteração de coloração. O Real-ESRGAN, por sua vez, se destaca por sua capacidade de preservar os contornos da imagem. Curiosamente, neste cenário, o SwinIR parece ter um resultado um pouco difuso, perdendo parte da composição estrutural. Já para o ruído Salt and Pepper, o SwinIR apresenta o melhor desempenho, sendo o mais eficaz na supressão desses artefatos. Apesar de ainda manter alguns resquícios, sua performance superior neste caso, aparentemente, revela o seu mecanismo de atenção, não percebendo esse ruído como parte de todo o contexto da imagem.



Figura 16: Imagem 0823 (DIV2K). Colunas: LR (*Motion Blur*), SwinIR, BSRGAN e Real-ESRGAN. Neste cenário de *deblurring*, os modelos SwinIR e BSRGAN apresentaram os resultados mais eficazes. O SwinIR, em particular, se destaca por produzir uma imagem visualmente mais nítida, mesmo que a reconstituição de alguns elementos pareça ainda um pouco "confusa", mesmo sendo o melhor resultado obtido. Já o Real-ESRGAN foi o modelo que menos corrigiu o borrimento.

Com essa análise, habilidades como a de tratar artefatos de compressão JPEG tornam estes modelos especialmente valiosos para aplicações gerais, visto que esta é uma condição fortemente prevalente nas imagens digitais. Também, a capacidade de lidar com múltiplos tipos de problemas, consequência direta de seu treinamento robusto, os torna relevantes como ferramentas de SR eficazes e versáteis.

VI. CONCLUSÃO

O projeto teve como objetivo realizar uma análise comparativa do desempenho de modelos de SR baseados em aprendizado profundo, a fim de avaliar sua robustez e eficácia sob diferentes cenários de degradação. A investigação se concentrou



Figura 17: Imagens WandaVision, Snow White (2025) (fancaps.net) e Helen Coates with the Bernstein children at Martha’s Vineyard, June 30, 1959 (loc.gov). As colunas seguem a ordem: Original, LR, SwinIR, BSRGAN e Real-ESRGAN.

A figura demonstra duas condições de degradação: as três primeiras linhas foram degradadas com Lanczos + Gaussian Blur, enquanto a última linha foi degradada apenas com o *downsampling* Lanczos. Com análise na reconstrução facial, o BSRGAN se mostra mais eficiente, mantendo uma melhor fidelidade anatômica, o que evidencia o impacto do dataset FFHQ em seu treinamento. Em contrapartida, este modelo tende a ser mais conservador na geração de texturas no restante da imagem. Já o SwinIR e o Real-ESRGAN, embora adicionem um nível maior de detalhes em outras áreas, produzem resultados menos eficazes nos rostos, que apresentam contornos e feições com aspecto atípico.



Figura 18: Imagens Portrait of Dave and Tom Cassinelli e View of center of Paterson from Garret Mountain (loc.gov). Colunas: Original, LR (*Nearest Neighbors*), SwinIR, BSRGAN e Real-ESRGAN. Com exceção do resultado ligeiramente mais suave do BSRGAN, os demais modelos produziram resultados visualmente muito similares, reconstruindo a imagem de forma eficaz.

em dois grupos de modelos: os chamados “clássicos”(EDSR, ESRGAN e SwinIR Clássica), treinados com a degradação bicúbica, e os de “super-resolução cega”(BSRGAN, Real-ESRGAN e uma versão de SwinIR Real), projetados para lidar com uma gama variada de deteriorações complexas de imagem.

A análise quantitativa revelou que, nas métricas de fidelidade como PSNR e SSIM, os modelos EDSR e SwinIR

apresentaram desempenho superior e consistente na maioria dos cenários. Em contrapartida, o ESRGAN se destacou na métrica perceptual LPIPS, confirmando seu foco em gerar resultados visualmente realistas, apesar de ser menos estável em degradações mais severas.

A avaliação subjetiva aprofundou esses achados, expondo um comportamento dual no ESRGAN: enquanto generalizava bem para borrões suaves, ele tendia a “alucinar”texturas irrealis ao ser exposto a artefatos de alta frequência, como os introduzidos pela interpolação de Hamming ou pelo ruído Salt and Pepper. Em contraste, o EDSR e o SwinIR mantiveram-se mais conservadores e fiéis à imagem de entrada, sendo que último demonstrou ser o mais completo, equilibrando alta fidelidade com uma reconstrução geométrica e de detalhes superior ao EDSR.

A análise dos modelos de super-resolução cega evidenciou sua clara superioridade em cenários de degradação complexos e realistas. Treinados com múltiplas deteriorações, o BSRGAN, o Real-ESRGAN e o SwinIR Real foram notavelmente mais eficazes em remover ruídos e corrigir borrões que os modelos clássicos não conseguiram tratar.

Conclui-se, portanto, que a escolha do modelo ideal de super-resolução depende criticamente da aplicação final e da natureza da degradação da imagem. Para cenários controlados onde a degradação é conhecida — especificamente interpolações como a bicúbica ou, surpreendentemente, a de Lanczos —, o SwinIR oferece o melhor balanço entre fidelidade e qualidade visual. Contudo, para aplicações do mundo real, onde as imagens sofrem com múltiplos problemas imprevisíveis como compressão JPEG, ruídos e borrões, os modelos de super-resolução cega são a abordagem mais recomendada, embora seja fundamental analisar tanto seus datasets de treinamento quanto a degradação a ser corrigida, com destaque para a versão do SwinIR, que apresentou o melhor desempenho em reverter borrões e ruídos severos.

REFERÊNCIAS

- [1] LEPCHA, Dawa Chyophel; GOYAL, Bhawna; DOGRA, Ayush; GOYAL, Vishal. **Image super-resolution: a comprehensive review, recent trends, challenges and applications**. Information Fusion, [S.I.], v. 91, p. 230–260, mar. 2023. Elsevier BV. Disponível em: <http://dx.doi.org/10.1016/j.inffus.2022.10.007>. Acesso em: 18 maio 2025.
- [2] LIM, Bee; SON, Sanghyun; KIM, Heewon; NAH, Seungjun; LEE, Kyoung Mu. **Enhanced Deep Residual Networks for Single Image Super-Resolution**. arXiv preprint arXiv:1707.02921, 2017. Disponível em: <https://arxiv.org/pdf/1707.02921.pdf>. Acesso em: 22 maio 2025.
- [3] LEDIG, Christian et al. **Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network**. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. p. 4681–4690. Disponível em: <https://arxiv.org/abs/1609.04802>. Acesso em: 28 maio 2025.
- [4] WANG, Xintao; YU, Ke; WU, Shixiang; GU, Jinjin; LIU, Yihao; DONG, Chao; LOY, Chen Change; QIAO, Yu; TANG, Xiaou. **ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks**. arXiv preprint arXiv:1809.00219, 2018. Disponível em: <https://arxiv.org/pdf/1809.00219.pdf>. Acesso em: 22 maio 2025.
- [5] WANG, Xintao et al. **Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data**. arXiv preprint arXiv:2107.10833, 2021. Disponível em: <https://arxiv.org/pdf/2107.10833.pdf>. Acesso em: 22 maio 2025.

- [6] ZHANG, Kai et al. Designing a Practical Degradation Model for Deep Blind Image Super-Resolution. arXiv preprint arXiv:2103.14006, 2021. Disponível em: <https://arxiv.org/pdf/2103.14006.pdf>. Acesso em: 22 maio 2025.
- [7] LIANG, Jingyun; CAO, Jiehang; SUN, Guolei; ZHANG, Kai; VAN GOOL, Luc; TIMOFTE, Radu. **SwinIR: Image Restoration Using Swin Transformer**. arXiv preprint arXiv:2108.10257, 2021. Disponível em: <https://arxiv.org/pdf/2108.10257.pdf>. Acesso em: 22 maio 2025.
- [8] LIBRARY OF CONGRESS. **Library of Congress Online Catalog**. Washington, DC: Library of Congress, 2025. Disponível em: <https://www.loc.gov/>. Acesso em: 27 maio 2025.
- [9] FANCAPS. **FanCaps: Movie and TV Screencaps Database**. 2025. Disponível em: <https://www.fancaps.net/>. Acesso em: 27 maio 2025.